DOCUMENT RESUME

ED 462 397 TM 025 061

AUTHOR Breunig, Nancy A.

TITLE A Review of Methods for Detection of Test and Item Bias.

PUB DATE 1996-01-00

NOTE 33p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25-27, 1996).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Chi Square; *Evaluation Methods; Heuristics; Identification; Item Response Theory; *Regression (Statistics); *Selection; *Test Bias; Test Use

IDENTIFIERS Decision Theory; *Item Bias Detection

ABSTRACT

Few issues have provoked as much controversy as the methods for detecting item and test bias. A recent illustration of the controversy surrounding this issue could be seen in the emotional reactions to the publication of "The Bell Curve." This paper reviews methods of evaluating both item and test bias. Small heuristic data sets are provided to illustrate required calculations. Test bias is presented in the context of bias in selection, and models based on regression and decision-theoretic approaches are discussed. Item bias is discussed by reviewing latent trait, chi-square, and item difficulty techniques. There is no definitive method for fair selection or for detection of biased items. The test user needs to consider the purpose of the test and the inferences that will be made from the test and decide based on these considerations the methods or combination of methods that would be best to use. In particular, with regard to item bias, it may be best to use a combination of methods and see if results from the different approaches are in agreement. (Contains 3 tables, 4 figures, and 15 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 462 397

Running head: METHODS FOR DETECTION OF TEST AND ITEM BIAS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
NANCY A. BREUNIG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

A Review of Methods for Detection of Test and Item Bias

Nancy A. Breunig

Texas A&M University

BEST COPY AVAILABLE

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, January, 1996.

TM025061

Abstract

Few issues have provoked as much controversy as the methods for detecting item and test bias. A recent illustration of the controversy surrounding this issue could be seen in the emotional reactions to the publication of The Bell Curve. The present paper will review methods of evaluating both item and test bias. Small heuristic data sets will be provided to illustrate required calculations. Test bias will be presented within the context of bias in selection. Models based on regression and decision-theoretic approaches will be discussed. Item bias will be discussed by reviewing latent trait, chi-square, and item difficulty techniques.

A Review of Methods for Detection of Test and Item Bias

Few issues have provoked as much controversy as the methods for detecting item and test bias. According to Reynolds and Kaiser (1990, p. 487), bias in testing has been a recurring controversy throughout the history of mental measurement. Much of this controversy has centered around the claim that traditional employment and educational tests are biased in favor of white middle class culture (Ironson & Subkoviak, 1979). A recent illustration of the controversy surrounding this issue could be seen in the emotional reactions to the publication of The Bell Curve.

Research on bias, which in the 1960's gained its own fervor surrounding reaction to the Civil Rights Movement, can be viewed as developing in two areas, bias in selection and item bias. Bias in selection is the study of bias in the presence of an external criterion, whereas, item bias is the study of bias in the absence of an external criterion (Ironson & Subkoviak, 1979).

The present paper will begin by defining bias and then review the various methods for detecting bias in selection and item bias. Small heuristic data sets will be employed to illustrate required calculations.

Definition of Bias

The majority of researchers define bias as invalid, systematic error in how an item or a test measures for members of a particular group (Camilli & Shepard, 1994; Shepard, 1982). Systematic measurement error is defined as error which

"...consistently affects an individuals's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured" (Crocker & Algina, 1986, p. 105).

Bias is systematic in the sense that it creates distortion in test results for members of a particular group. Tests do not perfectly measure an intended knowledge domain; therefore, as long as the measurement error affects scores for members of different groups equally, or the magnitude of random error is the same for all groups, a test or an item is not considered to be biased (Camilli & Shepard, 1994).

An example of a biased test can best be illustrated in the following scenario: two groups of examinees, African-American and Asian, are administered a math word-problem test. It appears from scores on the test that the African-American group scored higher than the Asian group. When the items are readministered to the two groups but presented orally instead of in written form, the differential difficulty disappears (Shepard, 1982, p. 11). Likewise, if the same two groups are taken to a track and timed on a quarter mile run, but a slower watch is used for the African-American group, it might appear that the Asian group is better at running. In actuality, no comparison can be made between groups because group averages would be confounded by bias in the methods of measurement. Only rankings within groups might be considered relatively accurate (Camilli & Shepard, 1994, p. 8).

The informed researcher will know that although group

differences in test performance exist, this does not always indicate bias (Camilli & Shepard, 1994). For example, in the math word-problem scenario, if the examiner intended to measure reading ability and not only math knowledge, the test may no longer be considered biased. According to Camilli and Shepard (1994), "...any conclusions made concerning test bias depend on the inferences drawn from and uses made of test results" (p. 9).

Bias in Selection

The following discussion focuses on test bias in terms of selection procedures or predictive/criterion-related validity and not construct validity. This is not to say that construct validity is not important or relevant. The construct validity of scores obtained from a test must have already been established before either bias in selection or item bias is investigated (Palomares & Friedrich, 1991).

Models based on regression

Tests are widely used in choosing applicants for educational or employment positions. Tests related to the criterion performance are used to identify the qualified applicants (Shepard, 1982, p. 15). Thus, the tests are being used for prediction and how accurately the tests identify qualified applicants is referred to as predictive validity. "Since the relationship between test and criterion performance is operationalized by regression equations, test bias was first defined as unequal regressions" (Shepard, p. 15). Cleary (1968), in her paper on the regression of college grades on the SAT for Negro and white students in integrated colleges, provides a

definition of bias in terms of regression. Cleary (1968) states,

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. (p. 115)

In Figure 1a, an example of an unbiased test regression is illustrated. For this example, even though the group means are different, the test has equal predictive validity for the Asian and African-American groups. If a member of the African-American group and a member of the Asian group had the same GRE (test) score, their expected graduate G.P.A. (criterion score) would also be the same.

In Figure 1b, an example of biased tests is illustrated. For this example, although the regression slopes are the same for both groups, the regression line for the Asian group is shifted to the right of the regression line for the African-American group. If a common regression line were used to predict the G.P.A. for the two groups, the GRE would underpredict performance of the Asian group and overpredict performance for the African-American group (Crocker & Algina, 1986).

There are a number of selection models based on determining separate regression lines for different groups and then calculating predicted criterion scores for each applicant by

using the appropriate regression coefficient determined by that candidate's group membership. Three such models, discussed in Crocker and Algina's (1986) textbook on test theory, are the Regression Model, the Equal-Risk (Employer's) Model, and Darlington's Model. Other models have been developed from different definitions of test bias. To discuss each model is beyond the scope of this paper. The interested reader is referred to Crocker and Algina (1986).

Models based on the Decision-Theoretic Approach

The remainder of the section on bias in selection will focus on the decision-theoretic approach. This model will be discussed at some length here, because it is generally preferred by psychometricians over other selection models. The reason this model is preferred is because decision-theoretic models force the test-user to establish a decision rule for selection of applicants based on a combination of the probabilities of events (true positive, true negative, false positive, and false negative) and their values to the decision maker (Crocker & Algina, 1986). In order to understand the decision-theoretic approach to selection, it is helpful to have an understanding of some basic terms and concepts.

Basics concepts. Figure 2a is a scatterplot of test scores (X) and criterion scores (Y) for a single group from a validation study. For this particular example, X will be GRE scores and Y will be graduate GPA. First, the examiner determines a point on the Y scale that will divide the examinees into successful and unsuccessful groups. In this example, as can be seen in Figure

2b, the examiner has chosen a G.P.A. of 2.0.

Next, the examiner sets a cutoff on X that will determine whether applicants are selected or rejected for admission into graduate school. In this example, Figure 2c, the examiner has chosen a GRE score of 900. In Figure 2d, we see that there are now four quadrants to our scatterplot. Quadrant A contains true positive events. A true positive event is when an examinee is selected that succeeds. Quadrant B contains false positive events. A false positive event is when an applicant is selected who subsequently fails. Quadrant C contains false negative events. A false negative events is when an applicant is rejected that could have succeeded. Quadrant D contains true negative events. True negative events are when an applicant is rejected that would have failed (Crocker & Algina, 1986).

Three other concepts are derived from this scatterplot. The other concepts are (a) the base rate, (b) the selection ratio, and (c) the success ratio. The base rate is the proportion of examinees who could succeed over the total number of examinees (A + D)/(A + B + C+ D). The selection ratio is the proportion of examinees who will be chosen (A + B)/(A + B + C+ D). The success ratio is the proportion of those examinees that are chosen who succeed A /(A + B) (Crocker & Algina, 1986; Nunnally & Bernstein, 1994).

Extensive-forms Analysis. There are two methods to the decision-theoretic approach to selection. The two methods are extensive-forms analysis and normal-forms analysis. The two methods are similar in their conceptual underpinnings; therefore,

only extensive-forms analysis will be discussed (Crocker & Algina, 1986).

According to Crocker and Algina (1986), there are four basic steps in extensive-forms analysis:

1. The probability of success and failure of an examinee is calculated.
2. The decision maker must assign weights, called utilities, to each possible result (success and failure) of a selection decision.
3. The expected utility of each decision alternative (select and reject) is calculated.
4. The expected utility values for each decision are compared and the decision with the greatest expected utility is chosen. (p. 278)

Prior to conducting extensive-forms analysis, the bivariate distribution of the predictor and criterion scores for all examinees has to have been determined through a previous validation study. The researcher also has to establish the criterion cutoff score dividing the examinees into successful and unsuccessful groups. Once this is completed, the researcher will be able to calculate the probabilities of success and failure for each examinee. The symbol $p|x$ will represent the probability of success given score x and $q|x$ will represent the probability of failure given score x (Crocker & Algina, 1986).

The researcher then decides which of the possible events (true positive, true negative, false positive, or false negative) are more important. Thus, the researcher assigns utility values

(weights) to each of the four possible events (Crocker & Algina, 1986). For example, say the researcher is trying to make a decision on whether to place a group of students in special education based on standardized test scores. For the researcher, the possibility of placing a student in special education when really does not belong there may be more important than not placing a student in special education who could benefit from these services. The researcher wants to avoid the stigmatization associated with labeling a child who would not benefit from special education services. Thus, the researcher decides:

$$U_{tp} = .75$$
$$U_{fp} = 1.00$$
$$U_{tn} = 1.00$$
$$U_{fn} = .25$$

In this example, the researcher has placed the most value or chose the highest utility for a false positive event (U_{fp}), placing a child in special education who will benefit from the services and the utility of a true negative event (U_{tn}), not placing a child in special education who would not benefit from the services. The researcher assigned these two events a utility of 1.00. The researcher has also placed the least value or utility (.25) on a false negative event U_{fn}, not placing a child in special education who would benefit from the services and a middle value of .75 for a true positive event U_{tp}, placing a child in special education who would benefit from the services. For more information on assigning numeric values to utilities, the reader is referred to Novick and Lindley (1978).

Now the researcher must calculate the expected utility of

each possible decision alternative and choose the alternative that has the highest expected utility value. The formula for calculating the expected utility for a select decision is:

$$E(u)_s = u_{sp}(p|X) + u_{fp}(q|X)$$

The formula for calculating the expected utility for a reject decision is:

$$E(u)_r = u_{fn}(p|X) + u_{sn}(q|X)$$

In our example, suppose p|x is .6 and q|x is .4. Thus,

$$E(u)_s = .75(.6) + 1(.4) = .85$$
$$E(u)_r = .25(.6) + 1(.4) = .55$$

The expected utility for a select decision exceeds the utility of a reject decision, so the researcher would select this particular student for special education services. The researcher would then perform this process for each student in the group to determine if they should receive special education services.

For this particular example, only one group of students was involved in the selection process. Since probabilities of success and failure may depend on group membership, the researcher may want to assign different utility values for different groups of students or applicants. With two groups of students/applicants, there are eight possible events. In our special education example, the researcher may want to calculate different expected utility values for students depending on whether they are non-native English speakers or native English speakers. The researcher may feel that it is more important that non-native English speakers not be placed in special education because the standardized test is in English and there is a chance the student

will score lower on the test as a result of not understanding the questions. The researcher decides that the utilities for the English speaking group will be: U_{tp} = .75, U_{fp} = 1.00, U_{tn} = 1.00, U_{fn} = .25. For the non-native English speakers the utilities will be: U_{tp} = .25, U_{fp} = 1.00, U_{tn} = 1.00, U_{fn} = .75.

If there are two examinees with the same predictor score and one is a native English speaker and the other is a non-native English speaker, then they may have different probabilities of success and failure. The probability of success for the English speaker will be .6 and the probability of failure will be .4. The probability of success for the non-native English speaker will be .4 and the probability of failure will be .6. Thus, for the English speaker the two expected utilities are:

$$E(u)_1 = .75(.6) + 1(.4) = .85$$
$$E(u)_2 = .25(.6) + 1(.4) = .55$$

For the non-native English speaker, the two expected utilities are:

$$E(u)_1 = .25(.4) + 1(.6) = .70$$
$$E(u)_2 = .75(.4) + 1(.6) = .90$$

Based on these calculations, the researcher would place the native English speaker in Special Education because the utility for the selection decision is greater than the utility for a reject decision. The researcher would not place the non-native English speaker in Special Education because the reject decision was greater than the select decision.

Item Bias

Once the construct validity of the scores from a test are established, an examiner may want to identify biased items before

construction of the final forms of the test. Detection of biased items is an *internal* method. The researcher is comparing an item to the total test and not to some outside criterion as in the models of fair selection. Although there are a number of different models for detecting item bias the researcher needs to be aware that all the models are based on information internal to the test itself (Scheuneman, 1979). Thus, they all assume that the average test item is unbiased, and will not detect item bias if there is constant bias across items (Ironson & Subkoviak, 1979). In other words, if that particular item has a high discrimination index, then the item is correlating well with the other items and it may be that all the items share some bias. In terms of analysis of variance, bias would be defined as item-by-group interaction (Cleary, 1968; Fisk, 1991).

A specific model for detecting item bias can indicate item bias when none exists or miss item bias that actually does exist. It is recommended that researchers not rely on the information yielded from just one method to determine if an item is biased, but instead use information from various methods (Crocker & Algina, 1986).

According to Crocker and Algina (1986):

> A set of items is *unbiased* if (1) the items are affected by the same sources of variance in both subpopulations; and (2) among examinees who are at the same level on the construct purportedly measured by the test, the distributions of irrelevant sources of variation are the same for

both subpopulations. (p. 377)

For example, if a math achievement test has both word problems and computational problems, the word problem items could be considered biased for non-native English speaking students. For this to be the case, the variance in the scores for the native English speakers and the non-native English speakers would not be due to the construct being measured (mathematical ability) but to proficiency in English. The variation in scores would not be affecting the two groups equally.

There are various methods for detecting item bias. The three most prominent methods in the literature are: (a) latent trait, (b) chi-square, and (c) item-difficulty (Crocker & Algina, 1986; Fisk, 1991). Each of these methods will be discussed briefly.

Latent trait or Item Response methods. Latent trait methods (Lawson, 1991; also referred to as Item Response methods), incorporate the use of item characteristic curves (ICC). The regression of item scores on ability level is called the item characteristic curve (Scheuneman, 1979). As shown in Figure 3a, an ICC portrays the relationship between an examinee's ability and the probability of the examinee answering the item correctly (Fisk, 1991; Ironson & Subkoviak, 1979).

The ICC can be described mathematically by an ability parameter denoted by θ and one or more item parameters. The items parameters are "a" the slope of the curve which is the item discrimination parameter, "b" the inflection point of the curve which is the item difficulty parameter, and "c" the lower asymptote which is the guessing parameter (Camilli & Shepard

1994; Fisk, 1991; Scheuneman, 1979). Figure 3b demonstrates an ICC with different a, b, and c parameters. For latent-trait models, an item is considered unbiased if the ICCs are the same for the different groups of interest (Ironson & Subkoviak, 1979).

There are two basic approaches to determining whether an item is biased. The first approach is to calculate the area between the ICCs for the two groups of interest, as a measure of the difference between the two ICCs. There are different formulas that can be used to calculate the area, such as Rudner's Area Measure. For more information on these different measures, the interested reader is referred to Crocker and Algina (1986). The second approach to detecting item bias is to perform statistical significance testing (Camilli & Shepard, 1994).

When calculating indices of item bias based on latent trait models, the researcher first assumes that the test from which the items are drawn is homogenous. That is, it is posited that the test measures one underlying trait or ability. Second, before the ICCs for an item for different groups can be compared, the item parameters must be scaled in the same metric. For a complete discussion on the two major methods for equating parameters, the interested reader is referred to Camilli and Shepard (1994). It should be noted that the guessing parameter is not affected by a change of scale, therefore, this parameter does not need to be equated or scaled in the same metric (Crocker & Algina, 1986).

Third, the researcher determines which model will be used to compare the ICCs. There are three models based on Item Response Theory to choose from. They are the one-parameter model, the two-

parameter model, and the three-parameter models (Camilli & Shepard, 1994; Crocker & Algina, 1986; Fisk, 1991; Lawson, 1991). Lastly, the researcher decides on the method(s) for detecting item bias. The researcher either computes area statistics and/or engages in statistical significance testing. With statistical significance testing, the test statistics (chi-square statistics) are computed to determine if the null hypotheses should be rejected. If a null hypothesis for that item is rejected, the item may be considered biased (Crocker & Algina, 1986). For heuristic purposes, only the one-parameter model will be discussed here, however, the two-parameter and three-parameter models follow the same principles.

The one-parameter model, sometimes referred to as the Rasch model (Rasch, 1980), sets the item discrimination parameter (a) to a constant and only the item difficulty parameter (b) is allowed to vary. The null hypothesis would be: H_o: $b_{1g} = b_{2g}$. The test of the null hypothesis is conducted and a chi-square statistic is obtained. The items for which the null hypothesis is rejected are considered to be biased items (Crocker & Algina, 1986). The chi-square statistic can also be regarded as an indication of the amount of bias in the item (analogous to effect size in ANOVA) (Crocker & Algina, 1986).

Latent trait methods are usually preferred over other methods of item bias detection because indications of item bias should not occur solely due to differences in the ability distribution. A disadvantage with latent-trait methods is that

even when item bias does not exist, differences in ICCs can still occur. If the items are not unidimensional, which was the assumption made for using latent-trait methods, an item may appear biased when all it is indicating is multidimensionality (Crocker & Algina, 1986). For the example discussed earlier about the math achievement test, the word-problems on the test would probably have different ICCs. Thus, the researcher could infer that the word problem items are biased. In actuality, the items are measuring both math knowledge and English proficiency (multidimensionality).

Another disadvantage to latent-trait methods is the sample size and number of items required. The three-parameter model requires approximately 1,000 subjects and 30 items or 500 subjects and 60 items. For the two-parameter model, a minimum of 500 examinees per group is suggested and for the one-parameter model 200 examinees in each group are required (Crocker & Algina, 1986).

Chi-square techniques. One alternative to the latent-trait methods that does not require such large sample sizes is the chi-square technique. With the chi-square technique samples with 100 or 200 examinees in each group are sufficient. In addition, chi-square techniques are easier to compute than latent-trait methods.

According to Crocker and Algina (1986), chi-square techniques "...essentially define an item as unbiased if, within a group of examinees with scores in the same test score interval, the proportion of examinees responding correctly to the item is

the same for both subpopulations" (p. 383). There are a number of different chi-square techniques.

In general, chi-square techniques divide the observed score scale into several intervals (Crocker & Algina, 1986). Three to five intervals are customary. For each interval, (a) a minimum of 10 to 20 correct responses should be included (b) expected frequencies should be at least five, and (c) a minimum number of incorrect responses should be decided upon (based on the researcher's judgement) and included (Fisk, 1991). Within each interval the groups of interest are compared in terms of proportions responding correctly to an item. If the proportions vary across groups, the item is considered biased (Crocker & Algina, 1986).

Two examples of chi-square, using Camilli's statistic, are provided in Table 1. These examples were generated for heuristic purposes and would never actually be used in research because they do not include a minimum number of incorrect responses. The author chose these examples so the reader could see two extremes, a chi-square where item bias is definitely present and where there is no bias present. For both examples, the observed score scale was divided into three intervals. N_{1j} refers to the number of examinees in the first group and second groups with scores in the jth interval. The symbols O_{1j} and O_{2j}, refer to the number of examinees in the first and second group who had scores in the jth interval and answered the item correctly.

P_{1j} is the proportion of examinees in the first group and the

jth interval that answered the item correctly: $P_{1j} = O_{1j} / N_{1j}$. The quantity $P._j$ is the proportion of all examinees who scored in the jth interval and answered the item correctly: $P._j = (O_{1j} + O_{2j}) / (N_{1j} + N_{2j})$. To compute Camilli's statistic the following formula is used:

$$\chi_c^2 = \Sigma \frac{N_{1_j} N_{2_j} (P_{1_j} - P_{2_j})^2}{(N_{1_j} + N_{2_j}) P_{._j} (1 - P_{._j})} = \Sigma \chi_j^2$$

After Camilli's statistic is computed, it can then be compared to a chi-square distribution with J degrees of freedom. J equals the number of intervals. Thus, J for these two examples would be three. As can be seen in the first example in the table, the chi-square statistic would be 39.99999. If we compare this value to a chi-square distribution, with 3 degrees of freedom at the .05 level, the critical value would be approximately 7.815. 39.99999 exceeds this value and would indicate that the item is biased. For the second example, the chi-squared statistic would be zero, indicating no bias.

Item difficulty techniques. A number of approaches exist for the use of item difficulty techniques in item bias studies. All the approaches are based on one of two definitions for a set of unbiased items. The first definition states that a set of items is unbiased if the item difficulties for the two groups are perfectly correlated (all points in the scatterplot lie in a straight line). The second definition states a set of items is unbiased if the difficulty difference between the two groups is

the same for all items. If the criterion for the second definition is met, then the criterion for the first definition will also be met but the reverse will not necessarily be true (Crocker & Algina, 1986).

The most widely implemented item difficulty technique, based on the first definition for a set of unbiased items, is commonly referred to as the Delta Plot Method. The Delta Plot Method was developed by Angoff and Ford (1973). This method involves computing the p-value, which is the proportion of examinees getting the item correct, for each item separately for each group. Next, the p-value is converted into a normal deviate *Z*, using tables of the standardized normal distribution. Once the z-value is obtained, in order to eliminate negative z values, a delta value is calculated using the formula: $\Delta = 4z + 13$ (Ironson & Subkoviak, 1979).

According to Fisk (1991), the delta values can then be "...plotted as ordered pairs on a bivariate graph" (p. 13). Although the correlation for the scatterplot may be high (e.g., .96) there still may be points that lie off the line that would be considered biased item(s).

An example of a Delta Plot method is provided in Tables 2 and 3. Table 2 is an example of hypothetical raw data collected from a dichotomous test on two groups of examinees with 15 examinees in each group. Table 3 shows that the item difficulty value (P) was computed for each item in each group. In addition, the Z-values corresponding to the appropriate P scores are reported which are then converted into delta values. Figure 4

illustrates a delta plot of the delta values for group 1 on the x-axis and group 2 on the y-axis. There are two outliers (items 2 and 10) that might be considered biased items.

Conclusions

As can be seen throughout this paper, there are not only different definitions of bias, but also different methods for bias detection based on the varying definitions. In the end, there is not one definitive method for fair selection or for detection of biased items. The test user needs to consider the purpose of the test and the inferences that will be made from the test and decide based on these considerations which methods or combination of methods would be best to use. Particularly, in regards to item bias, it may be in the best interest of the test user to employ a combination of methods and see if the results from the different methods are in agreement before deciding that an item is biased.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. CA: Brooks/Cole.

Angoff, W. H., & Ford, S. F. (1973). Item race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased items. Thousand Oaks, CA: Sage.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth: Harcourt Brace Jovanovich College.

Fisk, Y. H. (1991, January). A Brief Overview of Three Classes of Methods for Detecting Item Bias. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), (1991). Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

Novick, M. R., & Lindley, D. V. (1974). The use of more

realistic utility functions in educational applications. Journal of Educational Measurement, 15, 181-191.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Palomares, R. S., & Friedrich, K. R. (1991, January). Conceptual Underpinnings and Historical Perspectives on Evaluating Test Bias. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), The handbook of school psychology (pp. 487-521). New York: John Wiley & Sons.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152).

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore: John Hopkins University Press.

Table 1

Examples of Chi-Square using Camilli's Statistic

EXAMPLE 1

intervals	Score Level	N_{1j}	O_{1j}	P_{1j}	N_{2j}	O_{2j}	P_{2j}	$P_{.j}$		
1	8 - 10		20	10	0.5	20	20	1	0.75	13.33333
2	5 - 8		20	10	0.5	20	20	1	0.75	13.33333
3	1 - 5		20	10	0.5	20	20	1	0.75	13.33333

EXAMPLE 2

intervals	Score Level	N_{1j}	O_{1j}	P_{1j}	N_{2j}	O_{2j}	P_{2j}	$P_{.j}$		
1	8 - 10		20	10	0.5	20	10	0.5	0.5	0
2	5 - 8		20	10	0.5	20	10	0.5	0.5	0
3	1 - 5		20	10	0.5	20	10	0.5	0.5	0

BEST COPY AVAILABLE

Table 2

Hypothetical Data Set for the Delta Plot Method

Item	1	2	3	4	5	6	7	8	9	10
Group 1										
1	1	1	1	0	1	0	0	1	1	1
2	1	0	1	1	1	0	0	1	0	0
3	0	1	1	0	1	0	0	1	0	0
4	1	0	1	1	1	1	0	0	0	0
5	0	1	1	1	0	1	0	1	0	1
6	0	1	0	0	1	0	0	1	0	1
7	1	1	0	1	0	0	1	1	0	1
8	0	1	0	0	1	0	1	1	0	0
9	1	1	0	1	0	0	1	1	0	1
10	1	1	0	1	0	0	0	1	0	0
11	0	0	0	1	0	0	0	1	0	0
12	0	1	0	1	1	0	0	1	0	0
13	0	1	0	1	0	0	0	1	0	0
14	0	1	0	1	1	0	0	1	0	0
15	0	1	0	1	0	0	0	1	0	0
Group 2										
1	1	1	1	1	0	1	0	0	0	1
2	1	1	1	0	1	0	1	1	0	1
3	0	0	0	1	1	0	1	1	0	1
4	1	0	0	1	1	0	0	1	0	1
5	1	1	0	0	0	0	0	1	0	0
6	1	1	0	0	0	0	0	1	0	0
7	1	1	1	1	1	0	0	1	0	0
8	1	0	0	0	1	0	0	1	0	1
9	0	0	1	1	0	1	0	1	1	0
10	0	1	1	1	0	0	0	1	0	1
11	0	1	0	1	1	0	0	1	0	1
12	0	1	0	1	1	0	0	1	0	1
13	0	1	0	1	1	0	0	1	0	1
14	0	1	0	1	0	0	0	1	0	1
15	0	0	0	1	0	0	0	1	0	1

Table 3

Calculations for the Delta Plot Method

Item	P Score Group 1	P Score Group 2	Z Score Group 1	Z Score Group 2	Delta Group 1	Delta Group 2
1	0.4	0.46	-0.25	-0.1	12	12.6
2	0.8	0.66	0.85	0.4	16.4	14.6
3	0.33	0.33	-0.45	-0.45	11.2	11.2
4	0.73	0.73	0.6	0.6	15.4	15.4
5	53	0.53	0.05	0.05	13.2	13.2
6	0.13	0.13	-1.15	-1.15	8.4	8.4
7	0.2	0.13	-0.85	-1.15	9.6	8.4
8	0.93	0.93	1.5	1.5	19	19
9	0.06	0.06	-1.55	-1.55	6.8	6.8
10	0.33	0.73	-0.45	0.6	11.2	15.4

Figure Captions

Figure 1a. Unbiased tests according to the Cleary definition.

Figure 1b. Biased tests according to the Cleary definition.

Figure 2a. Scatterplot of test scores (x) and criterion scores (y).

Figure 2b. Scatterplot with the criterion cutoff.

Figure 2c. Scatterplot with predictor test cutoff.

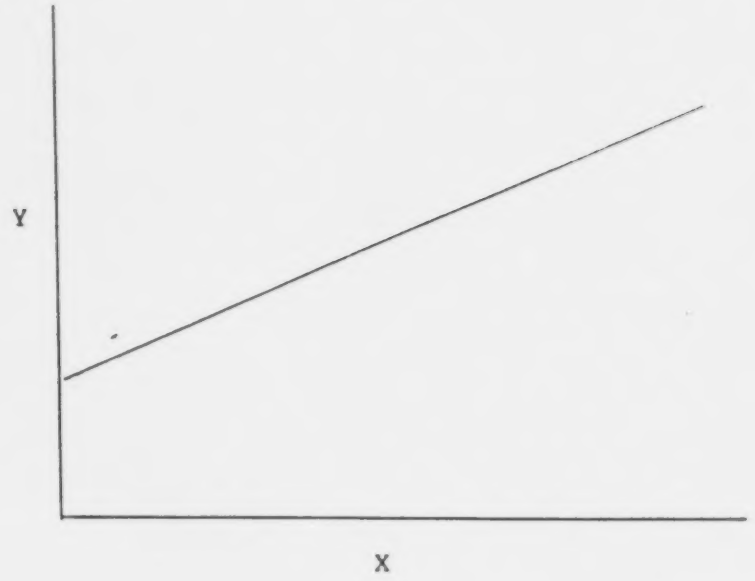
Figure 2d. Scatterplot with four quadrants.

Figure 3a. Item characteristic curve.

Figure 3b. ICC of an item that differs for both groups on the a, b, and c parameters.

Figure 4. Delta plot.

Figure 1a
TOP
Figure 1b
TOP



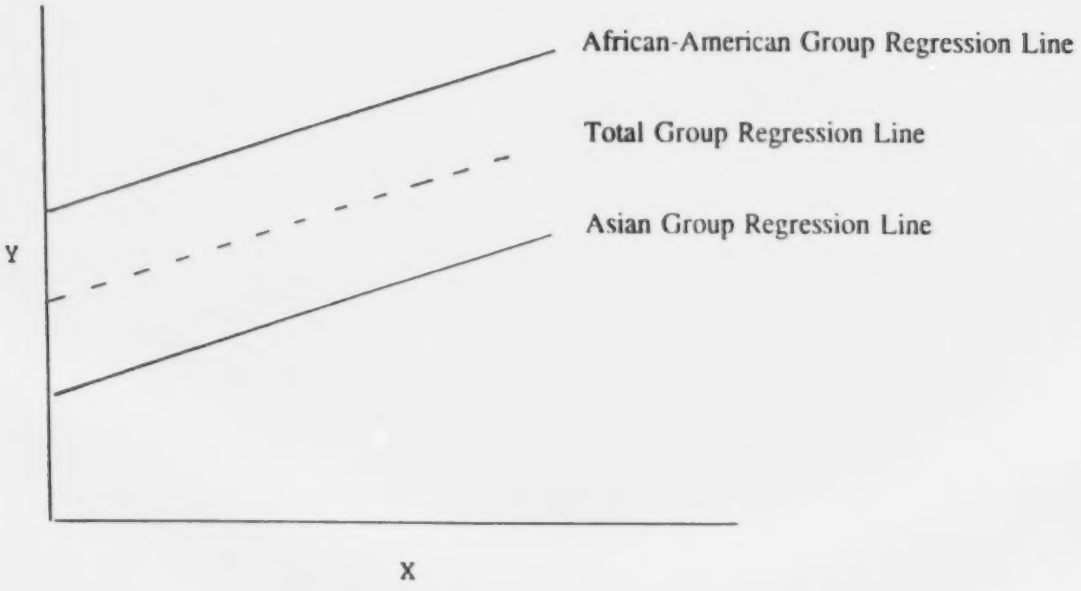


Figure 2a
TOP
Figure 2b
TOP

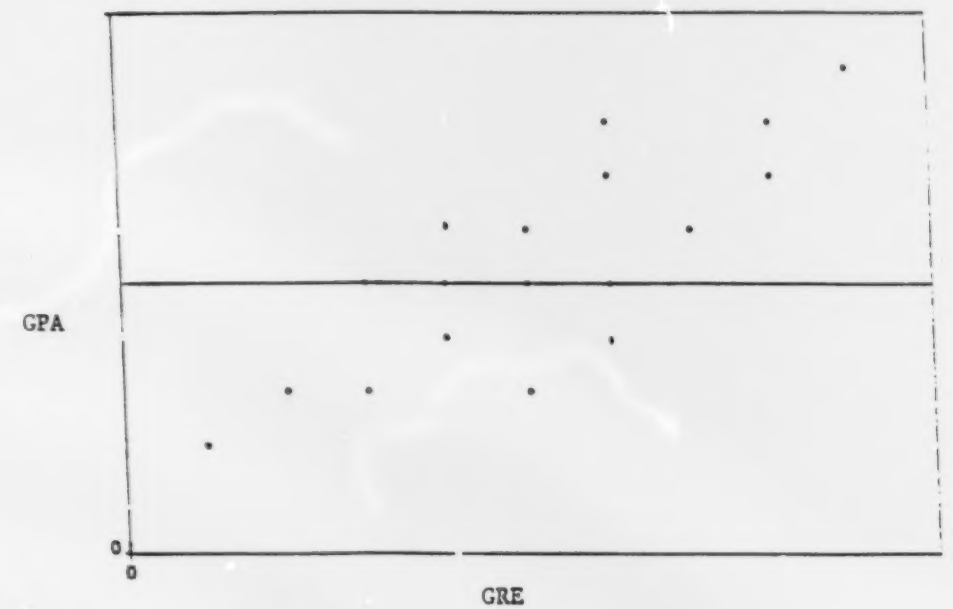
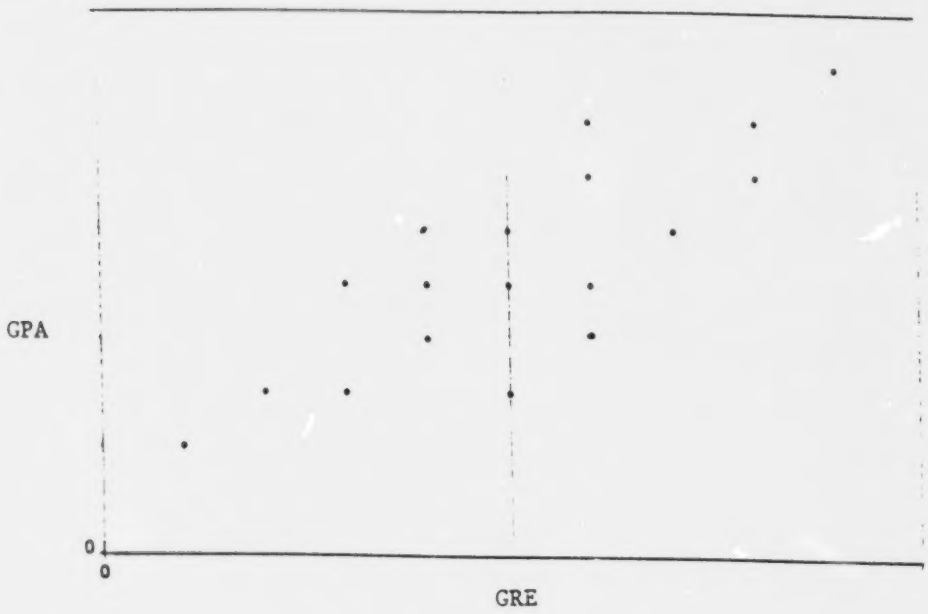


Figure 2c
TOP
Figure 2d
TOP



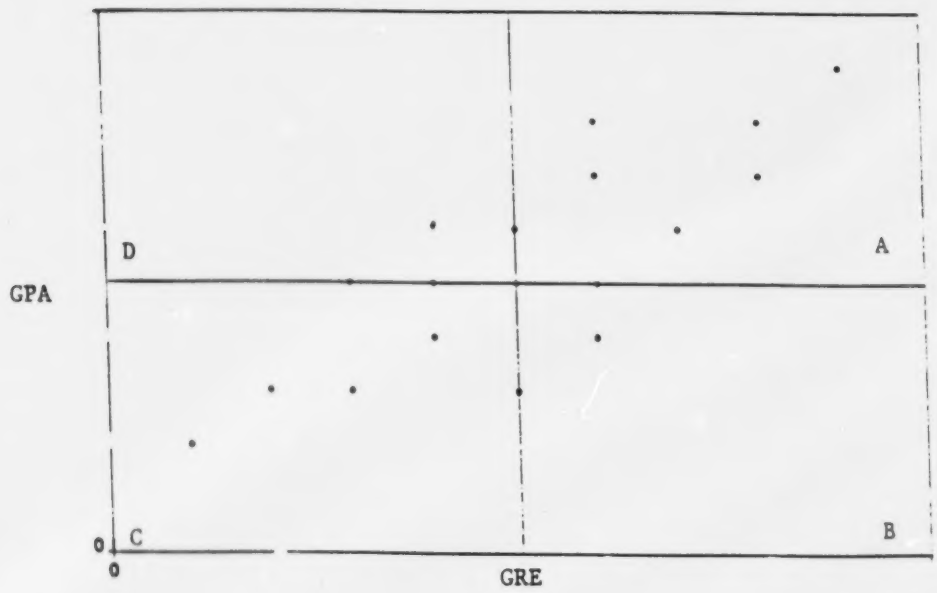
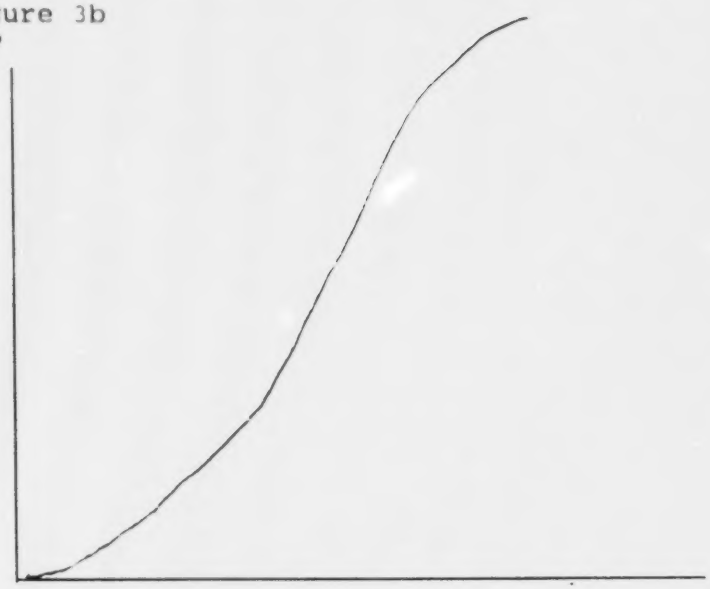


Figure 3a
TOP
Figure 3b
TOP



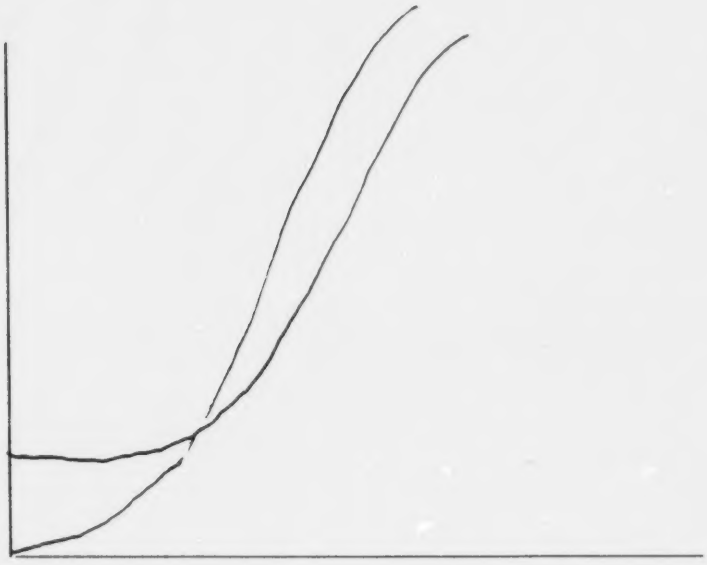


Figure 4
TOP